Sequential Exposure to Pemetrexed Followed by Cisplatin

Figure 4 shows isobolograms of the A549, MCF7, PA1, and WiDr cells exposed first to pemetrexed for 24 h and then cisplatin for 24 h. For the MCF7 cells, combined data points fell in the area of supra-additivity. The mean values of the observed data (0.40) were smaller than those of the predicted minimum values (0.44) (Table 1). The difference between them was significant ($p < 0.01$), indicating synergistic effects. For the A549 and PA1 cells, combined data points fell in the area of supra-additivity and within the envelope of additivity. The mean values of the observed data were smaller than those of the predicted minimum values (Table 1), but the differences were not significant ($p > 0.05$ and $p > 0.05$), indicating additive/synergistic effects. For the WiDr cells, the combined data points fell within the envelope of additivity and in the areas of supra-additivity and protection. The mean value of the observed data was smaller than the predicted maximum values and larger than that of the predicted minimum values (Table 1), indicating additive effects.

Sequential Exposure to Cisplatin Followed by Pemetrexed

Figure 5 shows isobolograms of the four cell lines exposed first to cisplatin for 24 h and then pemetrexed for 24 h. For the A549, MCF7, and PA1 cells, all or most of the combined data points fell in the areas of subadditivity and protection. The mean values of the observed data were larger than those of the predicted maximum values (Table 1). The differences were significant ($p < 0.05$, $p < 0.02$, and $p < 0.02$, respectively), indicating antagonistic effects. For the WiDr cells, most of the combined data points fell within the envelope of additivity, indicating an additive effect of this schedule.

Flow Cytometric Analysis

Finally, we evaluated the cytotoxic effects of pemetrexed and cisplatin on cancer cells using flow cytome-

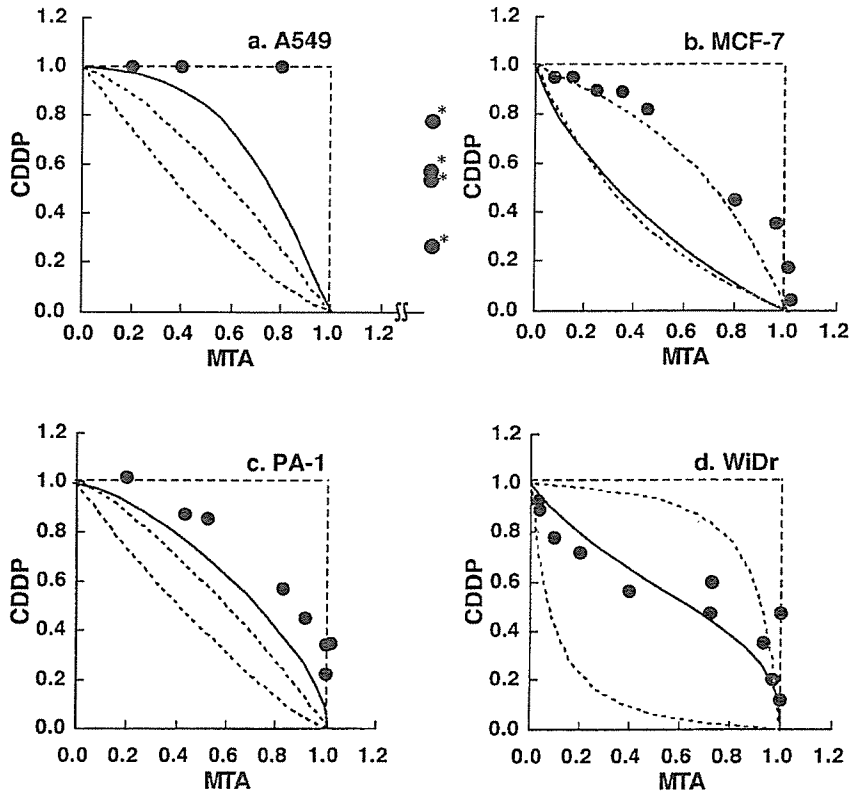

Figure 5. Isobolograms of sequential exposure to cisplatin (24 h) followed by pemetrexed (24 h) in A549 (a), MCF7 (b), PA1 (c), and WiDr (d) cells. For the A549, MCF7, and PA1 cells, all or most of the data points of the combinations fell in the areas of subadditivity and protection. For the WiDr cells, most of the data points of the combinations fell within the envelope of additivity and in the area of subadditivity. Data are mean values for at least three independent experiments; SE was <20%(*except the data).

Table 1. Mean Values of Observed, Predicted Minimum, and Predicted Maximum Data of Pemetrexed (MTA) in Combination With Cisplatin (CDDP) at IC_{80} for MCF7, PA1, and WiDr Cells and at IC_{50} for A549 Cells

Schedule	Cell Line	n	Observed Data	Predictaed Data for an Additive Effect		Effect
				Minimum	Maximum	
MTA + CDDP	A549	6	1.15	0.44	0.75	antagonism ($p < 0.05$)
	MCF7	8	0.95	0.57	0.72	antagonism ($p < 0.02$)
	PA1	9	0.69	0.40	0.56	antagonism ($p < 0.01$)
	WiDr	9	0.66	0.27	0.73	additive
MTA → CDDP	A549+	6	0.45	0.47	0.72	additive/synergism ($p > 0.05$)
	MCF7	9	0.40	0.44	0.78	synergism ($p < 0.01$)
	PA1	8	0.52	0.55	0.64	additive/synergism($p > 0.05$)
	WiDr	15	0.64	0.46	0.84	additive
CDDP → MTA	A549	7	1.14	0.41	0.74	antagonism ($p < 0.05$)
	MCF7	9	0.82	0.52	0.73	antagonism ($p < 0.02$)
	PA1	8	0.75	0.41	0.63	antagonism ($p < 0.02$)
	WiDr	11	0.71	0.21	0.82	additive

try. Cell cycle analysis revealed that pemetrexed and cisplatin arrested PA1 cells in late G_1 to early S phase and G_2/M phase, respectively (Fig. 6A, Table 2). When PA1 cells were exposed to both drugs simultaneously, the cell cycle profile was almost identical to that of a single treatment with pemetrexed, suggesting that the cell cycle effect of pemetrexed is dominant over that of cisplatin. As a result, the apoptosis-inducing effect of cisplatin, which was estimated by an increase in the size of sub-G_1 fraction, was almost completely cancelled in the presence of pemetrexed (Fig. 6A, MTA + CDDP). When PA1 cells were treated with cisplatin first and followed by pemetrexed, the cell cycle pattern closely resembled that of cells treated with cisplatin alone except for a modest increase in G_1 and S phases (Fig. 6A, Table 2, CDDP to MTA). The induction of apoptotsis was less prominent in the CDDP to MTA treatment than in the CDDP treatment (Table 2). In contrast, both apoptosis and G_2/M arrest were enhanced when PA1 cells were treated with pemetrexed first and followed by cisplatin compared with the treatment with either pemetrexed or cisplatin alone (Fig. 6A, Table 2, MTA to CDDP).

We carried out the same analysis with another cancer cell line MCF7 and obtained highly reproducible results. Upon simultaneous addition, the cell cycle effect of cisplatin was almost completely abrogated and the percentage of apoptotic cells was less than that of a single treatment with pemetrexed (Fig. 6B, MTA + CDDP). Similarly, apoptosis was suppressed when MCF7 cells were treated with cisplatin first and followed by pemetrexed compared with the treatment with either pemetrexed or cisplatin alone (Fig. 6B, Table 2, CDDP to MTA). In contrast, the apoptosis-inducing effect of pemetrexed was enhanced by the sequential exposure to cisplatin after pemetrexed (Fig. 6B, Table 2, MTA to CDDP). Overall, these data are fully consistent with the results of isobologram analysis, and provide the molecular basis of the interaction between the two drugs.

DISCUSSION

We found that the cytotoxic interaction between pemetrexed and cisplatin was schedule dependent. Simultaneous exposure to pemetrexed and cisplatin and sequential exposure to cisplatin followed by pemetrexed showed antagonistic effects in A549, MCF7, and PA1 cells, while sequential exposure to pemetrexed followed by cisplatin had a tendency to produce synergistic effects. In the latter schedule, observed data points in A549, MCF7, and PA1 cells were smaller than predicted minimum values for an additive effect (Table 1). WiDr cells showed additive effects in all schedules. The cause of difference in combined effects among cell lines is unknown. The difference may reflect the folate metabolism and the variety of target numbers (enzymes) in the cells. In addition, the isobologram of Steel and Peckham is stricter for synergism and antagonism than other methods. This may also influence the results.

In general, it is difficult to clarify the mechanisms underlying the drug combination. In this study, however, cell cycle analysis provided a clue to understand the molecular basis of schedule-dependent synergism and antagonism of the combination of pemetrexed and cisplatin. The exposure of PA1 and MCF7 cells to pemetrexed for 24 h led to a synchronization of most cells in late G_1 to

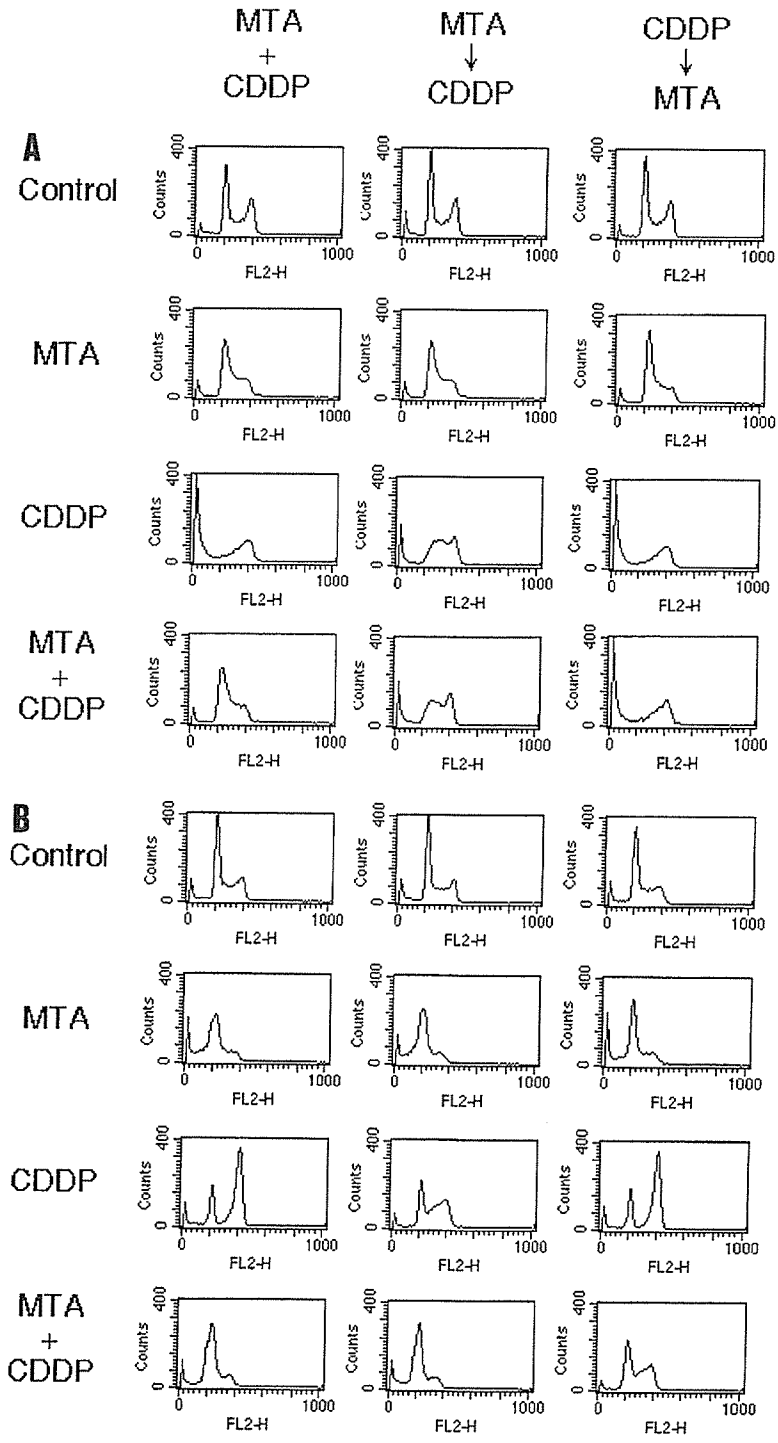

Figure 6. Flow cytometric analysis of cell cycle perturbation. PA1 cells, treated with 0.2 μM pemetrexed (MTA), 0.5 μM cisplatin (CDDP), both drugs simultaneously for 24 h, pemetrexed for 24 h followed by cisplatin for 24 h, or the reverse sequence were harvested at 48 h (A), and MCF7 cells, treated with 0.5 μM pemetrexed (MTA), 5 μM cisplatin (CDDP), both drugs simultaneously for 24 h, pemetrexed for 24 h followed by cisplatin for 24 h, or the reverse sequence were harvested at 48 h (B) and stained for DNA with propidium iodide and analyzed by flow cytometry as described in Materials and Methods.

Table 2. Cell Cycle Perturbations Induced by Pemetrexed (MTA), Cisplatin (CDDP), and Their Combinations for PA1 and MCF7 Cells at 48 h

	MTA + CDDP (24 h)				MTA (24 h) → CDDP (24 h)				CDDP (24 h)→ MTA (24 h)			
Cell Cycle (%)	Control	MTA	CDDP	MTA + CDDP	Control	MTA	CDDP	MTA + CDDP	Control	MTA	CDDP	MTA + CDDP
PA1 cells												
Sub-G_1	3.6	2.4	42.9	2.1	4.3	3.1	8.9	15.3	2.9	2.2	45.1	41.8
G_1	56.2	64.1	7.3	67.1	58.1	65.3	5.8	4.4	57.3	60.1	6.9	10.6
S	15.6	26.7	17.2	19.1	10.4	25.9	48.4	38.7	11.0	30.4	15.8	20.1
G_2/M	24.6	6.8	19.1	11.7	27.2	5.7	36.9	41.6	28.8	7.3	32.2	27.5
MCF-7 cells												
Sub-G_1	4.2	17.5	3.9	5.8	5.3	11.1	2.9	16.8	5.1	10.3	3.6	2.5
G_1	57.6	53.4	28.8	63.7	55.8	61.3	22.3	60.6	58.8	57.2	27.9	25.8
S	16.8	26.9	4.7	21.4	19.1	22.1	21.2	13.8	16.4	28.6	5.0	20.4
G_2/M	21.4	2.2	62.6	9.1	25.1	5.5	53.6	8.8	19.7	3.9	63.5	51.3

early S phase, in which cells are sensitive to cisplatin (20). This may explain the synergistic effects of sequential exposure to pemetrexed followed by cisplatin. On the contrary, one agent may reduce the cytotoxicity of the other agent by preventing cells from entering the specific phase in which the cells are most cytotoxic to the other agent. It has been shown that cisplatin elicits cytotoxic effects by blocking cells in G_2/M phase (20), while pemetrexed does by blocking cells in S phase (21). Indeed, simultaneous exposure to pemetrexed and cisplatin produced antagonistic effects, which were caused by the cancellation of cisplatin-induced G_2/M arrest by coexisting pemetrexed in PA1 and MCF7 cells. This was also the case with sequential exposure with cisplatin first followed by pemetrexed.

Our findings suggest that the sequential administration of pemetrexed followed by cisplatin may be the optimal schedule for these combinations. For example, administrations of pemetrexed on day 1 and cisplatin on day 2 would be worthy of clinical investigations. The simultaneous administration of pemetrexed and cisplatin and the sequential administration of cisplatin followed by pemetrexed may be inadequate. However, it must be noted that there are a number of difficulties in the translation of results from in vitro models to clinical therapy. The drug metabolism and pharmacokinetics under in vivo and in vitro conditions are different. Clinical outcome includes both the antitumor effects and normal tissue toxicity that results from a variable drug exposure, whereas in vitro models represent only antitumor effects at a constant drug exposure.

Teicher et al. studied the combination of pemetrexed with cisplatin in vivo against EMT-6 murine mammary carcinoma by a tumor cell survival assay (26). They observed that pemetrexed administered four times over 48 h with cisplatin administered with the third dose of pemetrexed produced an additive or more than additive tumor response. Teicher et al. further studied the combination of pemetrexed with cisplatin in human tumor xenografts (27). Administration of pemetrexed (days 7–11, days 14–18) along with cisplatin (day 7) produced greater-than-additive effects for human lung cancer H460 and Calu-6 tumor growth delay. Because experimental systems, schedules of drug administrations, and evaluating methods for synergism are different, it is difficult to compare their findings and ours.

A clinical and pharmacokinetic phase I study of pemetrexed in combination with cisplatin has been reported by Thordtmann et al. (15). They observed that this combination was clinically active and simultaneous administration of both agents on day on 1 (pemetrexed intravenously over 10 min and cisplatin over 2 h) every 21 days was less toxic than a sequential administration of pemetrexed on day 1 and cisplatin on day 2. They recommended the simultaneous administration of pemetrexed at 500 mg/m^2 plus cisplatin at 75 mg/m^2 on day 1 every 21 days for this combination. Phase II and III studies of the same schedules have been started for this combination and encouraging results have been obtained so far (16–18).

Our in vitro findings are not contradictory to clinical findings. In our study, simultaneous exposure to pemetrexed and cisplatin produced additive effects in WiDr cells and antagonistic effects in A549, MCF7, andPA1 cells. Most data points fell in the area of sudadditivity in MCF7 and PA1 cells, suggesting that the combination is superior to each drug alone but "sub-optimal." The simultaneous administration of pemetrexed and cisplatin was less toxic than the sequential administration, probably due to antagonistic interaction in the simultaneous exposure. Our isobologram shows that the doses of both agents in the pemetrexed–cisplatin sequence required

for IC_{80} or IC_{50} levels were much less (40–90%) than of those in simultaneous exposure (Fig. 3). Pemetrexed at 500 mg/m^2 and cisplatin at 75 mg/m^2, the optimal dose for the simultaneous administration, would be overdosed for the sequential administration of pemetrexed followed by cisplatin, which produced synergistic effects.

In conclusion, the present findings show that the interaction of pemetrexed and cisplatin is definitely schedule dependent. Sequential exposure to pemetrexed followed by cisplatin produced synergistic effects, whereas simultaneous exposure to the two agents and sequential exposure to cisplatin followed by pemetrexed produced antagonistic effects. These findings suggest that the optimal schedule of pemetrexed in combination with cisplatin at the cellular level is the sequential administration of pemetrexed followed by cisplatin. Although the simultaneous administration of pemetrexed and cisplatin on day 1 is more convenient and less toxic for patients than the sequential administration of pemetrexed on day 1 and cisplatin on day 2, the former schedule may be suboptimal and may not improve the clinical efficacy to "originally expected" level for this combination. It would be important to conduct dose-finding clinical trials in sequential administration of pemetrexed and cisplatin.

ACKNOWLEDGMENTS: This work was supported in part by a Grant-in-Aid for Cancer Research (11-8) from the Ministry of Health and Welfare and by a Grant-in-Aid for Research on the Second-Term Comprehensive 10-Year Strategy for Cancer Control from the Ministry of Health and Welfare of Japan.

REFERENCES

1. Taylor, E. C.; Kuhnt, D.; Shih, C.; Rinzel, S. M.; Grindey, G. B.; Barredo, J.; Jannatipour, M.; Moran, R. G. A dideazatetrahydrofolate analogue lacking a chiral center at C-6 N-[4-[2-(2-amino-34-dihydro-4-oxo-7H-pyrrolo[23-d]pyrimidin-5- yl)ethyl]benzoyl]-L-glutamic acid is an inhibitor of thymidylate synthase. J. Med. Chem. 35:4450–4454; 1992.
2. Habeck, L. L.; Mendelsohn, L. G.; Shih, C.; Taylor, E. C.; Colman, P. D.; Gossett, L. S.; Leitner, T. A.; Schultz, R. M.; Andis, S. L.; Moran, R. G. Substrate specificity of mammalian folypolyglutamate synthetase for 510-dideazatetrahydrofolate analogs. Mol. Pharmacol. 48: 326–333; 1995.
3. Shih, C.; Habeck, L. L.; Mendelsohn, L. G.; Chen, V. J.; Schultz, R. M. Multiple folate enzyme inhibition: Mechanism of a novel pyrrolopyrimidine-based antifolate LY231514 (MTA). Adv. Enzyme Regul. 38:135–152; 1998.
4. Shih, C.; Thornton, D. E. Preclinical pharmacology studies and the clinical development of a novelmultitargeted antifolate MTA (LY231514) In: Jackman, A. L., ed. Anticancer drug development guide: Antifolate drugs in cancer therapy. Totowa, NJ: Humana Press; 1998:183–201.
5. McDonald, A. C.; Vasey, P. A.; Adams, L.; Walling, J.; Woodworth, J. R.; Abrahams, T.; McCarthy, S.; Bailey, N. P.; Siddiqui, N.; Lind, M. J.; Calvert, A. H.; Twelves, C. J.; Cassidy, J.; Kaye, S. B. A phase I and pharmacokinetic study of LY231514 the multitargeted antifolate. Clin. Cancer Res. 4:605–610; 1998.
6. Rinaldi, D. A. Overview of phase I trials of multitargeted antifolate (MTA LY231514). Semin. Oncol. 26(Suppl. 6): 82–88; 1999.
7. Rusthoven, J. J.; Eisenhauer, E.; Butts, C.; Gregg, R.; Dancey, J.; Fisher, B.; Iglesias, J. Multitargeted antifolate LY231514 as first-line chemotherapy for patients with advanced non-small-cell lung cancer: A phase II study. National Cancer Institute of Canada Clinical Trials Group. J. Clin. Oncol. 17:1194–1199; 1999.
8. John, W.; Picus, J.; Blanke, C. D.; Clark, J. W.; Schulman, L. N.; Rowinsky, E. K.; Thornton, D. E.; Loehrer, P. J. Activity of multitargeted antifolate (pemetrexed disodium LY231514) in patients with advanced colorectal carcinoma: Results from a phase II study. Cancer 88:1807–1813; 2000.
9. Hanauske, A. R.; Chen, V.; Paoletti, P.; Niyikiza, C. Pemetrexed disodium: A novel antifolate clinically active against multiple solid tumors. Oncologist 6:363–373; 2001.
10. Pivot, X.; Raymond, E.; Laguerre, B.; Degardin, M.; Cals, L.; Armand, J. P.; Lefebvre, J. L.; Gedouin, D.; Ripoche, V.; Kayitalire, L.; Niyikiza, C.; Johnson, R.; Latz, J.; Schneider, M. Pemetrexed disodium in recurrent locally advanced or metastatic squamous cell carcinoma of the head and neck. Br. J. Cancer 85:649–655; 2001.
11. Shepherd, F. A. Pemetrexed in the treatment of non-small cell lung cancer. Semin. Oncol. 29(Suppl. 18):43–48; 2002.
12. Calvert, H. Pemetrexed (Alimta): A promising new agent for the treatment of breast cancer. Semin. Oncol. 30 (Suppl. 3):2–5; 2003.
13. Scagliotti, G. V.; Shin, D. M.; Kindler, H. L.; Scagliotti, G. V.; Shin, D. M.; Kindler, H. L.; Vasconcelles, M. J.; Keppler, U.; Manegold, C.; Burris, H.; Gatzemeier, U.; Blatter, J.; Symanowski, J. T.; Rusthoven, J. J. Phase II study of pemetrexed with and without folic acid and vitamin B12 as front-line therapy in malignant therapy in malignant pleural mesothelioma. J. Clin. Oncol. 21:1556–1561; 2003.
14. Hanna, N.; Shepherd, F. A.; Fossella, F. V.; Pereira, J. R.; De Marinis, F.; von Pawel, J.; Gatzemeier, U.; Tsao, T. C.; Pless, M.; Muller, T.; Lim, H. L.; Desch, C.; Szondy, K.; Gervais, R.; Shaharyar; Manegold, C.; Paul, S.; Paoletti, P.; Einhorn, L.; Bunn, Jr., P. A. Randomized phase III trial of pemetrexed versus docetaxel in patients with non-small-cell lung cancer previously treated with chemotherapy. J. Clin. Oncol. 22:1589–1597; 2004.
15. Thodtmann, R.; Depenbrock, H.; Dumez, H.; Blatter, J.; Johnson, R. D.; van Oosterom, A.; Hanauske, A. R. Clinical and pharmacokinetic phase I study of multitargeted antifolate (LY231514) in combination with cisplatin. J. Clin. Oncol. 17:3009–3016; 1999.
16. Manegold, C.; Gatzemeier, U.; von Pawel, J.; Pirker, R.; Malayeri, R.; Blatter, J.; Krejcy, K. Front-line treatment of advanced non-small-cell lung cancer with MTA (LY231514 pemetrexed disodium ALIMTA) and cisplatin: A multicenter phase II trial. Ann. Oncol. 11:435–440; 2000.
17. Shepherd, F. A.; Dancey, J.; Arnold, A.; Neville, A.; Rusthoven, J.; Johnson, R. D.; Fisher, B.; Eisenhauer, E. Phase II study of pemetrexed disodium a multitargeted antifolate and cisplatin as first-line therapy in patients with advanced non small cell lung carcinoma: A study

of the National Cancer Institute of Canada Clinical Trials Group. Cancer 92:595–600; 2001.
18. Vogelzang, N. J.; Rusthoven, J. J.; Symanowski, J.; Denham, C.; Kaukel, E.; Ruffie, P.; Gatzemeier, U.; Boyer, M.; Emri, S.; Manegold, C.; Niyikiza, C.; Paoletti, P. Phase III study of pemetrexed in combination with cisplatin versus cisplatin alone in patients with malignant pleural mesothelioma. J. Clin. Oncol. 21:2636–2644; 2003.
19. Scagliotti, G. V.; Kortsik, C.; Dark, G. G.; Price, A.; Manegold, C.; Rosell, R.; O'Brien, M.; Peterson, P. M.; Castellano, D.; Selvaggi, G.; Novello, S.; Blatter, J.; Kayitalire, L.; Crino, L.; Paz-Ares, L.; Go, R. S. Pemetrexed combined with oxaliplatin or carboplatin as first-line treatment in advanced non-small cell lung cancer: A multicenter, randomized, phase II trial. Clin. Cancer Res. 11(2 Pt. 1):690–696; 2005.
20. Adjei, A. A. Review of the comparative pharmacology and clinical activity of cisplatin and carboplatin. J. Clin. Oncol. 17:409–422; 1999.
21. Jackel, M.; Kopf-Maier, P. Influence of cisplatin on cell-cycle progression in xenografted human head and neck carcinomas. Cancer Chemother. Pharmacol. 27:464–471; 1991.
22. Tonkinson, J. L.; Marder, P.; Andis, S. L.; Schultz, R. M.; Gossett, L. S.; Shih, C.; Mendelsohn, L. G. Cell cycle effects of antifolate antimetabolites: Implications for cytotoxicity and cytostasis. Cancer Chemother. Pharmacol. 39: 521–531; 1997.
23. Tonkinson, J. L.; Worzalla, J. F.; Teng, C. H.; Mendelsohn, L. G. Cell cycle modulation by a multitargeted antifolate, LY231514, increases the cytotoxicity and antitumor activity of gemcitabine in HT29 colon carcinoma. Cancer Res. 59:3671–3676; 1999.
24. Schultz, R. M.; Dempsey, J. A. Sequence dependence of Alimta (LY231514, MTA) combined with doxorubicin in ZR-75-1 human breast carcinoma cells. Anticancer Res. 21:3209–3214; 2001.
25. Kano, Y.; Akutsu, M.; Tsunoda, S.; Izumi, T.; Mori, K.; Fujii, H.; Yazawa, Y.; Mano, H.; Furukawa, Y. Schedule-dependent synergism and antagonism between pemetrexed and paclitaxel in human carcinoma cell lines in vitro. Cancer Chemother. Pharmacol. 54:505–513; 2004.
26. Teicher, B. A.; Alvarez, E.; Liu, P.; Lu, K.; Menon, K.; Dempsey, J.; Schultz, R. M. MTA (LY231514) in combination treatment regimens using human tumor xenografts and the EMT-6 murine mammary carcinoma. Semin. Oncol. 28:55–62; 1999.
27. Teicher, B. A.; Chen, V.; Shih, C.; Menon, K.; Forler, P. A.; Phares, V. G.; Amsrud, T. Treatment regimens including themultitargeted antifolate LY231514 in human tumor xenografts. Clin. Cancer Res. 6:1016–1023; 2000.
28. Kano, Y.; Sakamoto, S.; Kasahara, T.; Akutsu, M.; Inoue, Y.; Miura, Y. In vitro effects of amsacrine in combination with other anticancer agents. Leukemia Res. 15:1059–1064; 1991.
29. Steel, G. G.; Peckham, M. J. Exploitable mechanisms in combined radiotherapy-chemotherapy: the concept of additivity. Int. J. Radiat. Oncol. Biol. Phys. 5:85–91; 1979.
30. Kano, Y.; Ohnuma, T.; Okano, T.; Holland, J. F. Effects of vincristine in combination with methotrexate and other antitumor agents in human acute lymphoblastic leukemia cells in culture. Cancer Res. 48:351–356; 1988.
31. Kano, Y.; Akutsu, M.; Tsunoda, S.; Mano, H.; Sato, Y.; Honma, Y.; Furukawa, Y. In vitro cytotoxic effects of a tyrosine kinase inhibitor STI571 in combination with commonly used antileukemic agents. Blood 97:1999–2007; 2001.
32. Kano, Y.; Akutsu, M.; Tsunoda, S.; Suzuki, K.; Adachi, K. In vitro schedule-dependent interaction between paclitaxel and SN-38 (the active metabolite of irinotecan) in human carcinoma cell lines. Cancer Chemother. Pharmacol. 42:91–98; 1998.
33. Furukawa, Y.; Iwase, S.; Kikuchi, J.; Nakamura, M.; Terui, Y.; Yamada, H.; Kano, Y.; Matsuda, M. Phosphorylation of bcl-2 protein by cdc2 Kinase during G2/M phases and its role in cell cycle regulation. J. Biol. Chem. 275:21661–21667; 2000.

Methods

Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis

Eugene Berezikov,[1] Geert van Tetering,[1] Mark Verheul,[1] Jose van de Belt,[1] Linda van Laake,[1,2] Joost Vos,[3] Robert Verloop,[3,4] Marc van de Wetering,[1] Victor Guryev,[1] Shuji Takada,[5] Anton Jan van Zonneveld,[3] Hiroyuki Mano,[5] Ronald Plasterk,[1,6] and Edwin Cuppen[1]

[1]*Hubrecht Laboratory, 3584 CT Utrecht, The Netherlands;* [2]*Department of Cardiology, HLCU Location, University Medical Center Utrecht, 3508 GA Utrecht, The Netherlands;* [3]*Department of Nephrology, Leiden University Medical Center, 2333 ZA Leiden, The Netherlands;* [4]*Department of Physiology, Free University Medical Center, 1081 BT Amsterdam, The Netherlands;* [5]*Division of Functional Genomics, Jichi Medical School, Kawachigun, Tochigi 329-0498, Japan*

MicroRNAs are 20- to 23-nucleotide RNA molecules that can regulate gene expression. Currently >400 microRNAs have been experimentally identified in mammalian genomes, whereas estimates go up to 1000 and beyond. Here we show that many more mammalian microRNAs exist. We discovered novel microRNA candidates using two approaches: testing of computationally predicted microRNAs by a modified microarray-based detection system, and cloning and sequencing of large numbers of small RNAs from different human and mouse tissues. Together these efforts experimentally identified 348 novel mouse and 81 novel human microRNA candidate genes. Most novel microRNAs candidates are not conserved beyond mammals, and ~10% are taxon-specific. Our analyses indicate that the entire microRNA repertoire is not remotely exhausted.

[Supplemental material is available online at www.genome.org. All novel microRNAs described in this study are submitted to the central miRNA repository miRBase, maintained at the Wellcome Trust Sanger Institute, under the ID nos. listed in the Supplemental material.]

Although the first microRNA (miRNA) was identified >10 yr ago (Lee et al. 1993; Wightman et al. 1993), it was only recently recognized that miRNAs form a major class of ribo-regulators (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001). Currently, >400 miRNAs have been described for human (Griffiths-Jones 2004), but estimates based upon computational predictions range between 500 and 1000 (Bentwich et al. 2005; Berezikov et al. 2005, 2006; Xie et al. 2005). miRNAs are transcribed as long precursors (pri-miRNAs) that are processed by Drosha, resulting in an ~70-nucleotide (nt) stem–loop structure (pre-miRNA), are transported to the cytoplasm, and are further processed by the Dicer-containing complex. The resulting 20- to 25-nt mature miRNAs are loaded in the RNA-induced silencing complex (RISC) that can effect gene silencing through sequence-specific base pairing with target mRNAs, resulting in either transcriptional repression or target breakdown (Bartel 2004; Du and Zamore 2005). Although only a limited number of biological targets has been validated experimentally, computational approaches, applying rules based on validated targets, suggest that the currently known set of miRNAs may regulate between 20% and 30% of all transcripts in a vertebrate genome (Lewis et al. 2005; Lim et al. 2005). Indeed, many developmental and disease processes have now been found to be under critical regulation by miRNAs (Alvarez-Garcia and Miska 2005; Plasterk 2006). To obtain better insight into the biological function of this class of small RNA molecules in general and individual miRNAs in particular, it will be essential to identify all miRNAs that are expressed from a genome. Here, we describe the identification of many novel miRNA candidates by following two different experimental approaches. First, we used a modified microarray-based detection system to experimentally confirm computationally predicted miRNAs; secondly, we generated several high-titer cDNA libraries for small RNAs from different human and mouse tissues, followed by sequencing and computational analysis of inserts. Together, these efforts identified 348 novel mouse and 81 novel human microRNA candidate genes and suggested that still more miRNAs exist in genomes.

[6]**Corresponding author.**
E-mail plasterk@niob.knaw.nl; fax 31-30-2516554.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5159906.

Results

Microarray-based confirmation of computationally predicted candidates

Previously, we reported computational predictions of >800 novel mammalian miRNA candidates that meet stringent phylogenetic conservation profiles and RNA folding criteria (Berezikov et al. 2005). To verify these predicted miRNA candidate genes experimentally, we now employed a modified RNA-primed Array-based Klenow Extension (RAKE) approach (Nelson et al. 2004). This assay is based on the ability of an RNA molecule to function as a primer for Klenow polymerase extension when fully base-paired with a single-stranded DNA molecule (Fig. 1A). As the exact 3′ end of the miRNA should be known for successful extension, and computational predictions are not optimal for predicting the correct start and end of the mature miRNA, we designed a tiling path of probes complementary to both known and predicted miRNA precursors (Fig. 1B). Such a tiling path RAKE assay is less prone to

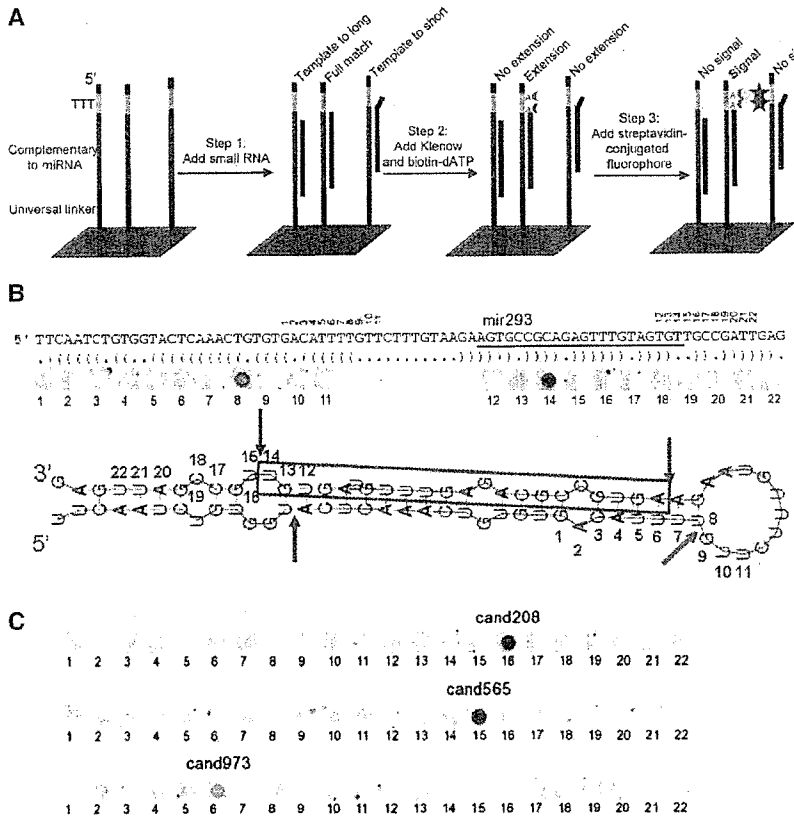

Figure 1. Schematic representation of the modified RAKE assay (*A*) and experimental results obtained for known (*B*) and novel (*C*) candidate miRNAs. (*A*) A 44K microarray with a tiling path of 60-mer probes that are attached with their 3′ end to the glass surface was designed for the Agilent platform. Each DNA probe consists of a universal 3′ spacer sequence (black) followed by 22 nt of sequence complementary to the microRNA candidate (red), three thymidine nucleotides (green), and a short universal spacer (black). Unlabeled small RNA is hybridized to these arrays, followed by a Klenow extension reaction in the presence of biotinylated dATP. As miRNAs function as a primer for extension and the thymidines are the only template for extension, a complete 3′-end match is required for biotin incorporation and streptavidin fluorophore-mediated detection in the final step. (*B*) Schematic representation of the miR-293 pre-miRNA with predicted secondary structure and RAKE results. The mature miRNA (red) and numbers *above* the sequence indicate the 3′ end that fully matches the respective tiling path probe on the array. The strongest signal in the RAKE assay is obtained for probe 14, confirming the known 3′ end of miR-293. A weaker signal is obtained for probe 8, consistent with sensitive detection of the star sequence, which is produced as a side product from the hairpin structure by Drosha and Dicer nucleases that cut double-stranded RNA with a 2-nt 3′ overhang. The star sequence cannot be detected for all positive miRNAs. (*C*) RAKE results for three candidate miRNAs confirm the existence of novel miRNAs and identify the 3′ end of the mature miRNAs. For cand208 the star sequence can be detected (probe 7), whereas for cand973 multiple probes are positive with a rapidly decreasing intensity around the predominant probe, most likely representing imprecise 3′ end processing.

false positives than standard hybridization assays, as it depends on the presence of a fully matching 3′ end of the miRNA and hence distinguishes between miRNA family members that differ in their 3′ sequences. Flanking tiling path probes function as negative controls.

Although some rules have been put forward to determine which strand of the stem is preferentially loaded as mature miRNA in the RISC complex (Khvorova et al. 2003; Schwarz et al. 2003), such computational predictions can be done only when the precise ends of the processed miRNA duplex are known. In addition, due to the nature of the hairpin sequence it is often difficult to predict which strand of the genomic DNA encodes a precursor. To take a fully unbiased approach, we designed tiling paths of 11 probes covering each arm of the stem–loop structure, for the sense as well as the antisense genomic sequence, resulting in sets of 44 probes per candidate microRNA gene. Due to G–U pairing allowed in RNA folding and different nucleotide composition of the complementary DNA strand, antisense transcripts do not necessarily fold into stable stem–loop structures, and for such candidates only 22 probes were included. The central position in the tiling path was determined by predicting the most likely Dicer/Drosha processing sites from secondary structure hairpin information.

We designed a custom validation microarray with 44,000 features, covering 259 known mouse miRNAs and 687 novel predicted miRNAs that are conserved between mouse and human, and filled up the array with 200 additional candidates based on stringent randfold criteria (Bonnet et al. 2004) and mouse and rat genome conservation. These arrays were probed with four different sources of small RNAs: mouse embryos at embryonic days 8.5 and 16.5, adult mouse brain, and embryonic stem (ES) cells (Fig. 1C). Mature miRNAs were semi-manually annotated after preprocessing the raw microarray output data using custom scripts. A nonredundant set of 184 of the known miRNAs (71%), 419 of the candidate conserved miRNAs (61%), and 128 of the extra set (64%) were found positive (Supplemental data 1, 2). The confirmed numbers of known and novel miRNAs may still be underestimates due to the limited number of different tissues assayed and the limited length of the tiling path used. Interestingly, for more than half of the known miRNAs the most prominent 3′ end observed in the RAKE assay differed from the annotated form, including several mature miRNAs residing in the other arm of the hairpin, suggesting that originally the star-sequence was annotated (Supplemental data 1). In addition, for various candidate and known miRNAs, multiple subsequent probes (2 or 3) resulted in a positive signal, indicating that 3′ end processing of miRNAs is not a completely accurate process at the single nucleotide level. These findings are in line with the observed variation in ends of cloned miRNAs (Aravin and Tuschl 2005).

Manual inspection of microarray signal along the tiling paths in known miRNAs suggests a high specificity of the RAKE assay, since we did not observe signal outside the expected locations (Fig. 1B). To directly estimate the rate of false-positive

discovery in RAKE, we tested a number of RAKE-positive candidates by Northern blot analysis (Supplemental Fig. 1). From a selected set of 29 candidates that had a clear signal on RAKE, 22 were found positive on Northerns (76%). For another set of 11 candidates that were classified as "inconclusive" based on RAKE results, 6 (55%) were positive on Northerns (Supplemental Table 1), suggesting that among candidates that failed in our RAKE experiments there is still a substantial fraction of real miRNAs. During the course of the experiments, 18 of the 28 miRNA candidates that were confirmed by Northern blot analysis were also identified by other research groups and are currently annotated in miRBase (Poy et al. 2004; Bentwich et al. 2005; Sewer et al. 2005; Xie et al. 2005; Mineno et al. 2006; Wheeler et al. 2006). However, as it will remain difficult to obtain solid numbers on false-negative rates, but more importantly also on false-positive rates of the RAKE assay, we propose to consider all RAKE-positive hairpins as candidate miRNAs until independent additional confirmation (e.g., cloning, Northern, PCR-based amplification) is obtained.

Small RNA cloning and analysis

The second approach we pursued is deep sequencing of size-fractionated small RNA libraries of isolated human and mouse tissues. Although it was suggested previously that such efforts had reached near saturation (Lim et al. 2003), only limited numbers of library clones from a selected set of vertebrate tissues have been sequenced. Moreover, our computational predictions and microarray-based confirmations suggested many novel miRNAs yet to be discovered. Therefore, we generated nine high-titer nonconcatamerized libraries of size-fractioned small RNAs from mouse brain and various human fetal tissues (brain, skin, heart, lung, and mixed tissues), primary cells (human umbilical cord blood-derived endothelial progenitor cells, UCB-EPC; human foreskin-derived microvascular endothelial cells, MVEC), and cell lines (mix of six colon tumor cell lines, CTCL) and sequenced in total >140,000 clones (Table 1). After vector and quality trimming, 87,704 inserts >18 bases were recovered.

We established a computational pipeline for automated annotation of the cloned sequences, taking into account unique chromosomal position, location in repetitive elements or rRNA, tRNA, snoRNA genes, conservation data from 18 animal genomes (human, chimp, macaque, mouse, rat, dog, cow, chicken, opossum, zebrafish, *fugu*, tetraodon, *Xenopus*, *Anopheles*, bee, fly, worm, *Ciona*), and secondary structure information using randfold (Bonnet et al. 2004). This analysis was applied to the mouse and human cloned fragments, as well as to all known human and mouse miRNAs and the positive candidates identified using the RAKE assay (Table 2; Supplemental Data 3). Two hundred twenty-two out of 335 mouse (66%) and 204 out of 328 human (62%) miRNAs, as deposited in miRBase (Griffiths-Jones 2004), passed the automated filtering and annotation. Most of the miRNAs that were discarded by the computational pipeline overlap with various repeat annotations. For the 873 mouse miRNA sequences that tested positive in RAKE, 312 were rejected ad hoc by the computational analysis for various reasons (Supplemental Data 1). To maintain very stringent criteria for all novel miRNA candidates reported here, we decided to exclude the latter set from further analyses and calculations, although they may represent genuine miRNAs.

For the sequenced small RNAs, 19,291 mouse (79%) and 23,351 human (60%) clones passed this filtering. Known abundant microRNA sequences dominate this set, but, interestingly, ~0.5% of the reads represent 54 novel mouse and 81 novel human miRNA genes (Supplemental Data 3). Underrepresentation of novel miRNAs in cloning and sequencing experiments can be expected, as there is a strong bias toward highly expressed and abundant miRNAs, which have already been identified in previous low-depth sequencing efforts. As a result, 76% of the novel miRNAs are supported by only a single clone, but it should be noted that 28% of the known miRNAs were also detected only once or not at all in our set of tissues (Fig. 2). It now becomes clear that expression levels of miRNAs may range over more than three orders of magnitude (Neely et al. 2006), which is indeed reflected in our cloning results where the most abundant miRNAs were picked up >3000 times and many others (both known and novel) only once (Supplemental data 4).

Overlap between approaches

Overlap between the novel miRNAs obtained by the RAKE-based computational verification experiments and the cloning experiment is limited (Fig. 3), potentially reflecting the fundamentally different biases underlying each approach, such as phylogenetic

Table 1. Small RNA library statistics

	Human brain	Human skin	Human heart	Human lung	Human mix1[a]	Human mix2[b]	Human UBC-EPC[c]	Human MVEC[d]	Human CTCL[e]	Mouse brain	Total
cfu	2.4×10^6	1.2×10^6	2.1×10^6	2.7×10^6	2.0×10^6	3.3×10^6	6.6×10^6	7.2×10^6	0.9×10^6	2.3×10^6	
PCR cycles[f]	20	17	20	17	22	22	17	17	18	15	
Total number of reads	11,520	4992	6144	6144	3840	8064	24,369	18,432	18,048	42,336	143,889
Accepted reads[g]	6761	4109	3949	4521	415	648	18,406	4878	12,429	31,588	87,704
Success rate	59%	82%	64%	74%	11%	8%	76%	26%	68%	75%	61%
Known miRNAs	157	164	121	144	29	50	144	121	154	206	
(reads)	(4745)	(2177)	(1380)	(2418)	(95)	(84)	(8328)	(2527)	(4886)	(19,407)	
Novel miRNAs	17	6	11	8	1	4	24	7	28	57	
(reads)	(21)	(8)	(11)	(12)	(1)	(4)	(34)	(13)	(51)	(155)	

[a]Various uncharacterized fetal tissues.
[b]Liver, stomach, bowel.
[c]Human umbilical cord blood-derived endothelial progenitor cells.
[d]Human foreskin-derived microvascular endothelial cells.
[e]Mix of six colon tumor cell lines.
[f]PCR cycles needed for cDNA amplification before cloning.
[g]After computational analysis.

Table 2. Computational analysis statistics for known and novel miRNAs

	Human				Mouse					
	All libraries		miRBase		Brain		RAKE		miRBase	
Category[a]	Reads	Loci	Reads	Loci	Reads	Loci	Reads	Loci	Reads	Loci
Accepted	56,116		328		31,588		873		335	
Mapped to genome	38,928	29,275	326	338	24,366	3869	870	2215	331	1573
Repeats	2868	24,588	52	62	1403	2124	68	1514	23	27
rRNA, tRNA etc.	3670	2856	5	4	1296	512	33	25	20	20
GC-rich, incorrect length	1781	57	6	6	796	9	12	9	3	4
Nonhairpin regions	3389	1165	10	7	1158	840	128	124	29	308
Known miRNAs within repeats	3263	42	42	50	210	21	7	20	15	906
Known miRNA passed randfold	23,196	193	204	199	19,136	178	220	169	222	223
Known miRNAs failed randfold	181	10	7	10	61	7	4	4	19	85
Hairpins passed randfold	155	81	0	0	155	57	341	296	0	0
Hairpins failed randfold	425	283	0	0	151	121	57	54	0	0

[a]See Methods for details.

conservation for the RAKE experiments and expression abundance in cloning from tissue samples. Furthermore, different tissue sources that potentially express different miRNAs were used in the various experiments to maximize the chance of finding novel miRNAs. Both the cloning approach and the microarray experiments show that the abundance of miRNAs varies over several orders of magnitude and that different tissues are characterized by a different group of highly expressed miRNAs (Supplemental data 4). Since the tissues used consist of many different cell types, miRNAs that are highly expressed in only a few cells of a tissue would seem to be expressed at low levels when the whole tissue is assayed and could easily be missed by cloning efforts. However, such miRNAs would be extremely powerful indicators for specific cell types (Wienholds et al. 2005) and, in line with previous reports, could be very valuable, for example, in classifying tumors (Calin et al. 2005; He et al. 2005; Lu et al. 2005).

To test the presence of RAKE-confirmed miRNAs in our small RNA libraries, we used information about the 3′ ends of miRNA to design primers that overlap with the miRNA sequence and the poly-A tail that was introduced during the cloning process. These oligos can be used in combination with a vector-based primer to amplify specific miRNA clones from total library DNA. From 32 known miRNAs tested, eight were positive in this PCR assay using the mouse brain library as template, indicating that in principle the assay can be used for specific amplification of miRNA clones, although the success rate is only 25% (Supplemental data 5). From 276 novel RAKE candidates that were tested by PCR, three candidates were positive (MM_25, MM_139, and MM_359). From these, two candidates (MM_25 and MM_139) were also observed in sequenced clones from mouse brain library, and MM_139 appeared to be a homolog of a recently published human miRNA, mir-551b (Cummins et al. 2006). Although confirmation rates are rather low, the results do also indicate that library sequencing is not exhausted and that more extensive sequencing of the libraries may yield more novel miRNAs and increase the overlap between miRNA candidates identified by RAKE and by cloning.

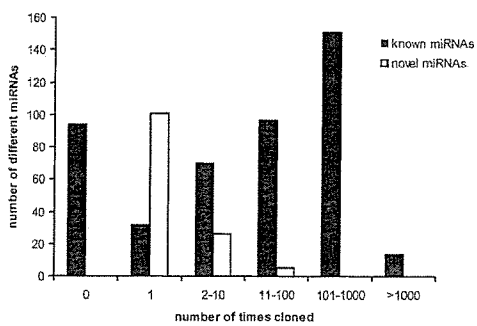

Figure 2. The cloning frequency for known and novel mouse and human miRNAs is significantly different. Although many known miRNAs were not picked up in our cloning experiments and about half of the novel miRNAs were identified multiple times, a relatively high fraction of novel miRNA candidates was identified only once, indicating that small RNA sequencing efforts have not been exhausted.

To further assess reliability of the RAKE assay, we compared mouse brain cloning and RAKE data. One hundred eleven known miRNAs were positive on RAKE using RNA from mouse brain and were also cloned from the mouse brain library. In 63 cases (57%), the most frequently cloned sequence had the same 3′ end as that identified in the RAKE experiment, and in 32 cases (29%), the miRNAs RAKE 3′ end corresponded to 3′ ends of some of the cloned inserts, but these were not the most frequently cloned sequences. Only in 16 cases (14%) were there no matching 3′ ends between RAKE and cloned sequences. This analysis demonstrates that, at least for known miRNAs, there is a good agreement between RAKE and cloning data, further substantiating the validity of the RAKE approach.

Discussion

According to the guidelines for miRNA annotation (Ambros et al. 2003), a cloned RNA must have either convincing evolutionary conservation or some other evidence of expression in order to be annotated as a genuine miRNA. Although most of the novel miRNAs presented here are supported by evolutionary conservation and stringent bioinformatic criteria, independent expression evidence is lacking for most of them. To distinguish between miRNAs that are extensively verified by independent experimental assays (e.g., Northern blots) and RNA species that are supported by limited single experiments (e.g., cloned only once, or only found positive in RAKE), but are highly likely to be real miRNAs, we prefer to use the term "miRNA candidate" for the latter class. Taken together, we identified 348 novel mouse (RAKE and cloning) and 81 novel human (cloning only) miRNA candi-

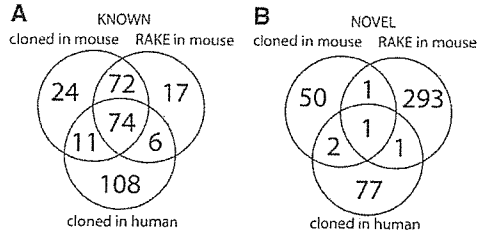

Figure 3. Overlap of cloned small miRNAs and RAKE results for known (*A*) and novel (*B*) mouse miRNA candidates. Known microRNAs were found to varying degrees in the different experiments, but most are detected in limited numbers of samples in both human and mouse. Most novel microRNA candidates were identified by the combination of computational prediction and RAKE confirmation approach.

date genes (Fig. 3). Novel miRNAs in general are expressed at a very low level and constitute <1% of clones in small RNA libraries. At the same time, novel miRNA candidates are less conserved than known miRNAs (Table 3). While the majority of known miRNAs are conserved among invertebrates, vertebrates, and mammals, only a few novel miRNAs are conserved to this extent. Instead, most of the novel miRNAs are conserved only in mammals, primates (for human-derived miRNAs), or rodents (for mouse-derived miRNAs). Seven percent of human and 11% of mouse novel miRNA are species-specific.

Similar to known miRNAs, several of the novel miRNA candidates were found to belong to families that have high sequence similarity, even though this characteristic has already been used to identify miRNAs by systematic genome analysis (Nam et al. 2005). In human, one family was identified that contains only novel miRNA candidates, and eight families were found to contain both known and novel miRNAs (Supplemental Table 2). In mouse, novel miRNA candidates make up 12 novel families and contribute to two known miRNA families.

Genomic clustering is a feature that has been employed for miRNA gene discovery previously (Altuvia et al. 2005), and from our newly identified set of miRNA candidates 19 mouse and 13 human miRNA genes reside in existing or novel (11 and five, respectively) genomic clusters (with 20 kb distance) of multiple miRNAs (Fig. 4; Supplemental Table 3). Interestingly, 10 miRNA candidates are novel members of a large cluster of >30 miRNAs.

Many known miRNAs reside in introns of protein-coding genes and in introns and exons of noncoding RNAs (Rodriguez et al. 2004). Similarly, 46% of novel human miRNA candidates are intronic (Table 4), and 27% overlap with both coding and noncoding exons, including 11% that overlap exon–intron junctions ("boundary" in Table 4). In mouse, 29%, 10%, and 13% of novel miRNAs are intronic, boundary, or exonic, respectively, and 48% of miRNAs do not overlap with known transcripts. The majority of novel miRNA candidates originate from the same strand as a host gene (66% in human and 73% in mouse).

Our results do indicate that miRNA discovery efforts are not exhausted. We have used very stringent parameters for the computational prediction of phylogenetically conserved miRNAs, thereby missing miRNAs that are less conserved or even species-specific. Indeed, using the RAKE assay, we could experimentally detect many novel miRNAs that were predicted without using stringent phylogenetic conservation criteria. Several miRNA candidates that gave inconclusive results on RAKE (and were excluded from further analysis) were still positive by Northern blot analysis, indicating that there are more real miRNAs among RAKE candidates than we were able to confirm. Furthermore, we discarded many RAKE-positive candidates that failed the computational pipeline developed for analysis of sequences from small RNA libraries. We decided to comply with these very stringent filtering criteria for consistency with the cloning approach, and to obtain the most reliable data set, although one should realize that we may have filtered out genuine miRNAs. Indeed, a number of known miRNAs did not pass the pipeline under the selected stringent parameters (Table 2), and several nonpassed RAKE candidates were confirmed by Northern blot analysis (Supplemental Table 1). The 296 miRNA RAKE candidate regions that passed the computational pipeline overlap poorly with our cloning efforts (Fig. 3), with only three miRNAs confirmed by cloning and one additional RAKE candidate recovered from the mouse brain library by directed PCR (Supplemental data 5). Positive results of directed PCR indicate that more RAKE candidates can be confirmed by increasing the sequencing depth of small RNA libraries. At the same time, the limited success rate of the PCR-based confirmation suggests that many RAKE candidates may not be present in the libraries. Since the confirmation rate of RAKE candidates by Northern blot analysis is >50% (Supplemental Fig. 1), it is possible that some real miRNAs are simply not "clonable" by the methods used, for example, due to their physical properties or modifications, and therefore cannot be detected in small RNA libraries. Although we provide strong bioinformatics and expression evidence for 296 RAKE miRNA candidates, for 290 of them we do not have experimental evidence for the exact location of the 5′ end of the mature miRNA sequence, and therefore these miRNAs should still be considered as candidate miRNAs.

Our cloning efforts also suggest that more miRNAs can be discovered in the future. Many known and novel miRNAs were represented by just a single clone in the various libraries (Fig. 2). Near saturation is expected only when the vast majority of miRNAs has been identified multiple times. Both increased depth analysis of existing libraries and cell-type-specific analyses are likely to further increase the number of miRNAs. To minimize the false-discovery rate, we have applied stringent criteria for computational analysis of small RNA sequencing reads that discarded several hundred potential miRNA hairpins along with a number of known miRNAs (Table 2). Future improvements in the bioinformatics analysis combined with experimental methods of independent verification of miRNA candidates may further increase the yield of novel miRNAs from small RNA sequencing efforts. Eventually, a comprehensive inventory of mammalian microRNA genes will facilitate elucidation of biological functions of this class of genes and aid in the identification of the involvement of individual microRNAs in development and disease processes (Alvarez-Garcia and Miska 2005; Hammond 2006; Plasterk 2006).

Table 3. Conservation of known and novel miRNAs

Conservation level	Human		Mouse	
	Known	Novel	Known	Novel
Invertebrates	28	2	24	2
Vertebrates	96	5	107	34
Mammals	61	27	53	221
Order-specific[a]	13	41	18	53
Nonconserved	1	6	2	38

[a]Primates for human, and rodents for mouse.

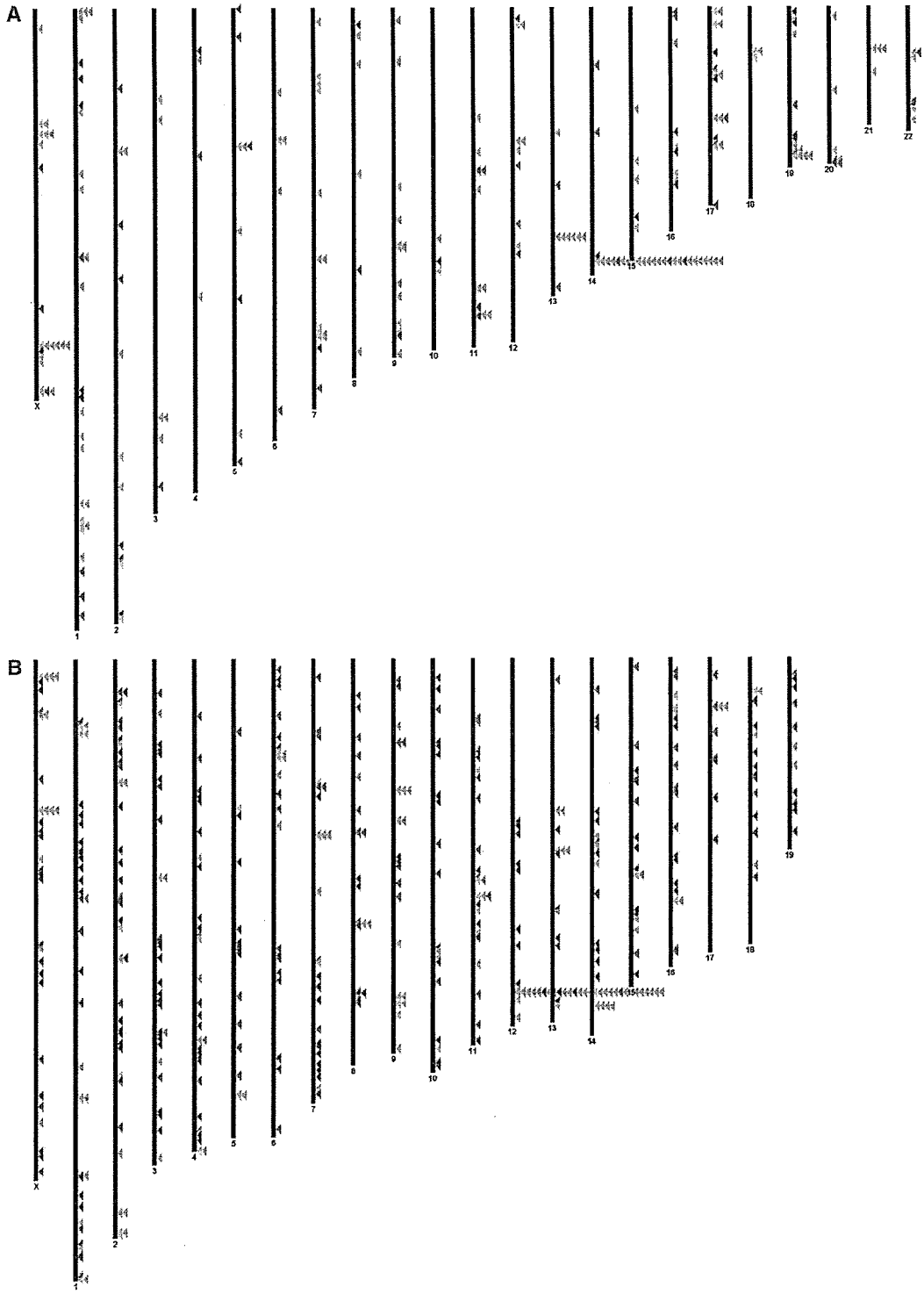

Figure 4. Chromosomal location of human (*A*) and mouse (*B*) miRNAs. (Red) Novel miRNA candidates, (green) known miRNAs. Clustered miRNAs are stacked.

Table 4. Genomic context of novel miRNAs

Strand	Location	Coding[b]	Human	Mouse
	Intergenic		22	168
+	Intronic	+	25	48
+	Intronic	−	9	16
−	Intronic	+	3	33
−	Intronic	−	0	4
+	Boundary[a]	+	6	12
+	Boundary	−	3	5
−	Boundary	+	0	10
−	Boundary	−	0	7
+	Exonic	+	4	17
+	Exonic	−	8	8
−	Exonic	+	1	13
−	Exonic	−	0	7

[a]"Boundary" is a case in which a miRNA hairpin overlaps an intron–exon junction.
[b]For intronic and boundary locations, noncoding means that surrounding exons are noncoding.

Methods

Computational prediction of novel microRNAs

All conserved human–mouse microRNAs were deduced from the candidate microRNA set as described in Berezikov et al. (2005). To produce additional candidate microRNA genes, the mouse genome was scanned for potential hairpins with a sliding window of 100 nt, and randfold values were calculated for resulting hairpins (mononucleotide shuffling, 1000 iterations). From a large set of hairpins that have low randfold values but are not necessarily conserved in other species, a subset of 200 was randomly selected to fill up the array.

RAKE microarray design and analysis

The microarray for verification of candidate microRNAs using the RAKE assay was designed as a 44K custom microarray (Agilent Technologies). Probes (60-mer) that are attached to the glass surface with their 3′ end were designed to include a fully matching probe sequence of 22 nt complementary to the predicted microRNA with universal spacers on each side (5′ end, 5′ spacer: CGATCTTT, sequence of 22 nt complementary to the microRNA candidate region [tiling path], 3′ spacer: TAGGGTCCGATAAGG GTCAGTGCTCGCTCTA, 3′ end attached to glass surface). The three Ts in the 5′ spacer function as a template for Klenow-mediated microRNA extension using biotin-dATP. A tiling path of 11 nt was designed to cover the most likely Dicer/Drosha cleavage site determined at 22 nt upstream and downstream from the terminal loop extended to contain at least 11 unpaired nucleotides. For all cases, probes were designed for both arms of the hairpin sequence, and for 648 candidates an additional set of 2 × 11 probes was designed, as the transcript originating from the antisense genomic sequence can also efficiently fold into a stable hairpin structure. All 22/44 probes for a candidate microRNA were located in clusters on the array to exclude regional background effects. Ten different hybridization controls complementary to plant microRNAs (miR-402, UUCGAGGCCUAUU AAACCUCUG; miR-418, UAAUGUGAUGAUGAACUGACCU; miR-167, UGAAGCUGCCAGCAUGATCUGG; miR-416, GGUUCGUACGUACACUGUUCAU; miR-173, UUCGCUUGCA GAGAGAAAUCAC; miR-417, GAAGGUAGUGAAUUUGUUC GAC; miR-163, GAAGAGGACUUGGAACUUCGAU; miR-419, UUAUGAAUGCUGAGGAUGUUGU; miR-405, GAGUUGGGU CUAACCCAUAACU; miR-420, UAAACUAAUCACGGAAAUG CAC) were represented 10 times randomly distributed on the array. Microarrays were scanned on an Agilent scanner model G2565B at 10-μm resolution, and spot identification and intensity determination was done using Agilent Feature Extraction software (Image Analysis version A.7.5.1) with standard settings. To permit manual inspection and annotation of mature microRNA sequences, the raw images and spot intensity data were processed using custom scripts and visualized together with tiling path sequence information (Supplemental data 2). Web-based interfaces were designed for annotation of single experiments and for summarizing all experiments. After manual inspection, all novel mature microRNA sequences that were positive (Supplemental data 1) were fed into the bioinformatics analysis pipeline set up for the evaluation of the cloned small RNAs, to filter out signal originating from repetitive elements and structural RNAs and to find homologous miRNAs in other species (Table 2).

Modified RAKE assay

The original RAKE assay (Nelson et al. 2004) was modified for use with high-density custom-printed microarrays in the Agilent platform. Most importantly, in contrast to most custom-spotted microarrays, custom-printed probes are attached with their 3′ end to the glass surface. This excludes the need for the exonuclease step that was included in the original protocol to reduce background signal from fold-backs of the free 3′ ends of the probes that result in double-stranded DNA structures that can function as a template for the Klenow extension, resulting in nonspecific background signal. Furthermore, hybridization, washing, and incubation conditions were adapted. All hybridization and wash buffers were made fresh from autoclaved stock solutions using DEPC-treated water, filter-sterilized, and preheated. Microarray slides and coverslips were pre-washed two times for 2 min at 37°C with preheated wash buffer (2× SSPE, 0.025% N-lauroylsarcosine), followed by 5 min incubation with pre-hybridization buffer (5× SSPE, 40% formamide, 0.025% N-lauroylsarcosine). Next, the Agilent hybridization chamber was completely filled with hybridization mix, leaving no air bubbles, as the usual air bubble for mixing does not move around at low temperature and with the hybridization mix used. The hybridization mix (750 μL total per slide) consists of 500 μL of 1.5× hybridization buffer (7.5× SSPE, 60% formamide, 0.0375% N-lauroylsarcosine), 10 μL of spike-in RNA (control plant microRNAs stock: miR-402, 1×10^{-6} M; miR-418, 3.3×10^{-7} M; miR-167, 1×10^{-7} M; miR-416, 3.3×10^{-8} M; miR-173, 1×10^{-8} M; miR-417, 3.3×10^{-9} M; miR-163, 1×10^{-9} M; miR-419, 3.3×10^{-10} M; miR-405, 1×10^{-10} M; miR-420, 3.3×10^{-11} M), and 20 μg of small RNA sample (8.5 dpc and 16.5 dpc mouse embryo, mouse embryonic stem [ES] cells, and total brain), isolated using the MirVana microRNA isolation kit (Ambion) and supplemented with DEPC-treated water up to 240 μL. The hybridization mix was heated for 5 min to 75°C and cooled on ice before application to the array. The array was incubated overnight at 37°C, followed by four washes of 2 min in wash buffer and one wash for 2 min in 1× Klenow buffer (10 mM Tris pH 7.9, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, 0.025% N-lauroylsarcosine). For the Klenow extension, an enzyme mix (750 μL total per slide) containing 375 μL of 2× Klenow buffer, 365 μL of DEPC-treated water, 2.5 μL of Klenow Exo− (50,000 U/μL, NEB), and 7.5 μL of biotin-14-dATP (4 μM stock, Perkin Elmer) was applied to the array in a clean incubation chamber and incubated for 1 h at 37°C. Next, the array was washed four times for 2 min with wash buffer and once for 2 min with 1× Klenow buffer. Next, the dye conjugation mix (total volume 750 μL) consisting of 375 μL of 2× Klenow buffer, 368 μL of DEPC-

treated water, and 20 µL of streptavidin-conjugated Alexa fluor-647 (2 mg/mL stock, Invitrogen) was applied in a new incubation chamber for 30 min at 37°C, followed by four washes of 2 min at 37°C with wash buffer and five brief dips in DEPC water to remove salts. Slides were dried by centrifugation in a 50-mL tube by spinning for 5 min at 1000 rpm (180 × *g*).

Northern blot analysis

The small RNA fraction of 16.5 dpc mouse embryos was isolated using the mirVana miRNA isolation kit (Ambion). Two micrograms of small RNA was loaded per lane and separated on 12% denaturing polyacrylamide gels alongside the RNA Decade marker (Ambion), transferred by electroblotting to positively charged nylone membranes (Roche). Blots were hybridized overnight at 37°C with radioactively (^{32}P) labeled DNA oligo probes in modified Church and Gilbert buffer, washed three times with 2× SSC, 0.1% SDS at 37°C, and visualized using phosphoimaging (Typhoon).

Small RNA library construction and sequencing

Nine high-titer small RNA libraries (Table 1) were made by Vertis Biotechnology AG. For human fetal tissue collection, individual permissions using standard informed consent procedures and prior approval of the ethics committee of the University Medical Center Utrecht were obtained. Briefly, the small RNA fraction from adult mouse brain (12 wk), various human fetal tissues (17 wk of development: brain; heart; skin; lung; mix 1: multiple fetal tissues; mix 2: liver, stomach, bowel), UCB-EPC (primary human endothelial progenitor cells, cultured from CD34+ cells purified from human umbilical cord blood), MVEC (primary human foreskin-derived microvascular endothelial cells), and colon tumor cell lines (mix of six cell lines) were isolated using the mirVana microRNA isolation kit (Ambion), followed by an additional enrichment by excision of the 15- to 30-nt fraction from a polyacrylamide gel. For cDNA synthesis, the RNA molecules in this fraction were first poly A-tailed using yeast poly(A) polymerase followed by ligation of a RNA linker oligo to the 5′ phosphate of the miRNAs. First-strand cDNA synthesis was then performed using an oligo(dT)-linker primer and M-MLV-RNase H- reverse transcriptase. The resulting cDNA was then PCR amplified for 15–22 cycles (depending on the starting material quality and quantity; see Table 1 for details), followed by restriction nuclease treatment, gel purification of the 95- to 110-bp fraction, and cloning in the EcoRI and BamHI sites of the pBSII SK+ plasmid vector. Ligations were electroporated into T1 Phage-resistant TransforMaxTMEC100TM electrocompetent cells (Epicentre), resulting in titers between 1.2 and 3.3 × 10^6 recombinant clones per library. A total of 83,328 colonies was automatically picked into 384-well plates (Genetix QPix2) containing 75 µL of LB-Amp and grown overnight at 37°C with continuous shaking. All of the following pipetting steps were performed using liquid-handling robots (Tecan Genesis RSP200 with integrated TeMo96 and Velocity11 Vprep with BenchCell 4×). Five microliters of culture were transferred to a 384-well PCR plate (Greiner) containing 20 µL of water, and cells were lysed by heating for 15 min at 95°C in a PCR machine. One microliter of lysed suspension was transferred to a fresh 384-well plate containing 4 µL of PCR mix (final concentrations: 0.2 µM M13 forward, TGTAAAACGACGGC CAGT; 0.2 µM M13 reverse, AGGAAACAGCTATGACCAT; 400 µM of each dNTP; 25 mM tricine; 7.0% glycerol [w/v]; 1.6% DMSO [w/v]; 2 mM $MgCl_2$; 85 mM ammonium acetate pH 8.7; and 0.2 U Taq polymerase in a total volume of 10 µL), and the insert was amplified by 35 cycles of 20 sec at 94°C, 10 sec at 58°C, and 30 sec at 72°C. After adding 30 µL of water, 1 µL of PCR product was directly used for dideoxy sequencing by transferring to a new 384-well PCR plate containing 4 µL sequencing mix (0.027 µL of BigDye terminator mix v3.1 [Applied Biosystems], 1.96 µL of 2.5× dilution buffer [Applied Biosystems], 0.01 µL of sequencing oligo [100 µM stock T7, GTAATACGACTCAC TATAGGGC], and 2 µL of water). Thermocycling was performed for 35 cycles of 10 sec at 94°C, 10 sec at 50°C, 20 sec at 60°C, and final products were purified by ethanol precipitation in 384-well plates as recommended by the manufacturer (Applied Biosystems) and analyzed on ABI3730XL sequencers with a modified protocol for generating ~100-nt sequencing reads.

PCR-based miRNA verification of library DNA

About 0.6 × 10^6 cfu from the mouse brain small RNA library was used for inoculation of a 100-mL liquid culture (LB + ampicillin). After overnight growth, plasmid DNA was isolated using minicolumns according to the manufacturer's instructions (Qiagen). Ten nanograms of plasmid DNA was used as a template for PCR using a vector-based forward primer (M13 forward) and miRNA-specific reverse primers that are complementary to the last 14 nt of the candidate miRNA and 8 nt of the poly-A tail that was introduced in the cloning procedure. PCR and sequencing (using T7 primer) conditions were identical to those described above for the small RNA library sequencing.

Computational analysis of cloned small RNAs sequencing reads

Base calling and quality trimming of sequence chromatograms were done by Phred software (Ewing et al. 1998). After masking of vector and adapter sequences, and removing redundancy, inserts of length ≥18 bases were mapped to genomes (ncbi35 assembly for human and ncbim34 assembly for mouse) using MegaBlast software (ftp://ftp.ncbi.nlm.nih.gov/blast/). Not all inserts matched perfectly to a genome, and detailed analysis of nonmatching sequences indicated that in most cases the best hit in the genome corresponded to the beginning of the read, with several nucleotides at the 3′ end of the read nonmatching (Supplemental Fig. 2A). The fraction of known miRNAs in perfect and nonperfect reads is similar (Supplemental Fig. 2B), justifying trimming of the nonmatching bases. These nongenomic sequences may be artifacts of the cloning procedure or a result of nontemplated modification of mature microRNAs (Aravin and Tuschl 2005). Such sequences were trimmed at the 3′ end according to the best blast hit to a genome. Next, for every genomic locus matching to an insert, repeat annotations were retrieved from the Ensembl database (http://www.ensembl.org) and tRNA, rRNA, snoRNA regions, and repetitive regions were discarded from further analysis, with the exception of simple and trf repeats, since these repeat annotations overlap with some known miRNAs. Genomic regions containing inserts with 100-nt flanks were retrieved from the Ensembl and RNAshapes programs (Steffen et al. 2006), and were used to find hairpin structures in sliding windows of 80, 100, and 120 nt that folded into hairpins with the abstract shape "[ ]," had a probability of folding >0.8, and contained an insert in one of the hairpin arms.

To find homologous hairpins in other genomes, mature miRNA regions were blasted against human, chimpanzee, macaque, mouse, rat, dog, cow, opossum, chicken, zebrafish, *fugu*, tetraodon, *Xenopus*, *Anopheles*, *Drosophila*, bee and *Ciona* genomes. Where available, BLASTZ_NET aligned regions were also retrieved from Ensembl. All hits matching to at least seven continuous nucleotides starting from the 1st, 2nd, or 3rd nucleotide of the mature sequence were extracted and folded using the RNAshapes program with the same parameters as mentioned

above. Next, similarity between all potential homologous hairpins and the original hairpin was calculated using RNAforester software (http://bibiserv.techfak.uni-bielefeld.de/rnaforester). If a BLASTZ_NET aligned region folded into a hairpin and had an RNAforester score >0.3, it was assigned as an orthologous hairpin in a particular species; otherwise, the highest scoring hairpin above a score of 0.3 was defined as an ortholog. Next, homologs from different organisms were aligned with the original hairpin by CLUSTAL W (Thompson et al. 1994) to produce a final multiple alignment of the hairpin region. Chromosomal locations of homologous sequences were used to retrieve gene and repeat annotations from the respective species in the Ensembl database. Hairpins that contained repeat/RNA annotations in one of the species, as well as hairpins containing mature regions >25 nt or with GC content >85% were discarded. For the remaining hairpins, randfold values were calculated for every sequence in an alignment using mononucleotide shuffling and 1000 iterations (Bonnet et al. 2004). The cut-off of 0.005 was used for randfold, and only regions that contained a hairpin below this cut-off for at least one species in an alignment were considered as miRNA candidates. Finally, positive hairpins were split into known and novel miRNAs according to annotations. To facilitate these annotations and also to track performance of the pipeline, mature sequences of known microRNAs from miRBase v.8.0 (Griffiths-Jones 2004) were included into the analysis.

Acknowledgments

We thank Wigard Kloosterman and Rene Ketting for critically reading the manuscript. This work was supported by grants from the Horizon (E.B.) and BioRange (E.C.) programs of the Netherlands Genomics Initiative (NGI). J.V. was supported by a grant from the Dutch Program for Tissue Engineering. R.V. was supported by a grant from the Netherlands Heart Foundations (NHS 2002157).

References

Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M.J., Tuschl, T., and Margalit, H. 2005. Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res.* **33:** 2697–2706.

Alvarez-Garcia, I. and Miska, E.A. 2005. MicroRNA functions in animal development and human disease. *Development* **132:** 4653–4662.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., et al. 2003. A uniform system for microRNA annotation. *RNA* **9:** 277–279.

Aravin, A. and Tuschl, T. 2005. Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett.* **579:** 5830–5840.

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* **37:** 766–770.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H., and Cuppen, E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120:** 21–24.

Berezikov, E., Cuppen, E., and Plasterk, R.H.A. 2006. Approaches to microRNA discovery. *Nat. Genet.* **38:** S2–S7.

Bonnet, E., Wuyts, J., Rouze, P., and Van de Peer, Y. 2004. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20:** 2911–2917.

Calin, G.A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S.E., Iorio, M.V., Visone, R., Sever, N.I., Fabbri, M., et al. 2005. A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N. Engl. J. Med.* **353:** 1793–1801.

Cummins, J.M., He, Y., Leary, R.J., Pagliarini, R., Diaz Jr., L.A., Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A.E., Labourier, E., et al. 2006. The colorectal microRNAome. *Proc. Natl. Acad. Sci.* **103:** 3687–3692.

Du, T. and Zamore, P.D. 2005. microPrimer: The biogenesis and function of microRNA. *Development* **132:** 4645–4652.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Griffiths-Jones, S. 2004. The microRNA registry. *Nucleic Acids Res.* **32:** D109–D111.

Hammond, S.M. 2006. MicroRNAs as oncogenes. *Curr. Opin. Genet. Dev.* **16:** 4–9.

He, L., Thomson, J.M., Hemann, M.T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S.W., Hannon, G.J., et al. 2005. A microRNA polycistron as a potential human oncogene. *Nature* **435:** 828–833.

Khvorova, A., Reynolds, A., and Jayasena, S.D. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115:** 209–216.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853–858.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858–862.

Lee, R.C. and Ambros, V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862–864.

Lee, R.C., Feinbaum, R.L., and Ambros, V. 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75:** 843–854.

Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. 2003. Vertebrate microRNA genes. *Science* **299:** 1540.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769–773.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., et al. 2005. MicroRNA expression profiles classify human cancers. *Nature* **435:** 834–838.

Mineno, J., Okamoto, S., Ando, T., Sato, M., Chono, H., Izu, H., Takayama, M., Asada, K., Mirochnitchenko, O., Inouye, M., et al. 2006. The expression profile of microRNAs in mouse embryos. *Nucleic Acids Res.* **34:** 1765–1771.

Nam, J.W., Shin, K.R., Han, J., Lee, Y., Kim, V.N., and Zhang, B.T. 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.* **33:** 3570–3581.

Neely, L.A., Patel, S., Garver, J., Gallo, M., Hackett, M., McLaughlin, S., Nadel, M., Harris, J., Gullans, S., and Rooke, J. 2006. A single-molecule method for the quantitation of microRNA gene expression. *Nat. Methods* **3:** 41–46.

Nelson, P.T., Baldwin, D.A., Scearce, L.M., Oberholtzer, J.C., Tobias, J.W., and Mourelatos, Z. 2004. Microarray-based, high-throughput gene expression profiling of microRNAs. *Nat. Methods* **1:** 155–161.

Plasterk, R.H.A. 2006. MicroRNAs in animal development. *Cell* **124:** 877–881.

Poy, M.N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P.E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P., et al. 2004. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* **432:** 226–230.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., and Bradley, A. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **14:** 1902–1910.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115:** 199–208.

Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M.J., Tuschl, T., van Nimwegen, E., and Zavolan, M. 2005. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics* **6:** 267.

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R. 2006. RNAshapes: An integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22:** 500–503.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Wheeler, G., Ntounia-Fousara, S., Granda, B., Rathjen, T., and Dalmay, T. 2006. Identification of new central nervous system specific mouse

microRNAs. *FEBS Lett.* **580:** 2195–2200.
Wienholds, E., Kloosterman, W.P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H.R., Kauppinen, S., and Plasterk, R.H.A. 2005. MicroRNA expression in zebrafish embryonic development. *Science* **309:** 310–311.
Wightman, B., Ha, I., and Ruvkun, G. 1993. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75:** 855–862.
Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.

Received January 20, 2006; accepted in revised form July 10, 2006.

Leukemia & Lymphoma, January 2006; 47(1): 9–14

Taylor & Francis
Taylor & Francis Group

REVIEW

DNA micro-array analysis of myelodysplastic syndrome

HIROYUKI MANO

Division of Functional Genomics, Jichi Medical School, Kawachigun, Tochigi, Japan

Abstract
Myelodysplastic syndrome (MDS) is an enigmatic disorder characterized by ineffective hematopoiesis and dysplastic morphology of blood cells. The clinical course of MDS consists of distinct stages, with early stages often progressing to advanced ones or to acute myeloid leukemia (AML). Little is known of the molecular pathogenesis of MDS or of the mechanism of its stage progression. DNA micro-array analysis, which allows simultaneous monitoring of the expression levels of tens of thousands of genes, has the potential to provide insight into the pathophysiology of MDS. Several studies have applied this new technology to compare gene expression profiles either between MDS and the healthy condition, among the different stages of MDS or between MDS-derived AML and *de novo* AML. Selection of an appropriate hematopoietic fraction is important for such studies, which to date have been performed with differentiated granulocytes, $CD34^+$ progenitors and $CD133^+$ immature cells. These studies have revealed that each stage of MDS has its own 'molecular signature', indicating the feasibility of differential diagnosis of MDS based on gene expression profile. They have also demonstrated that the current clinical diagnosis of MDS results in the misclassification of patients with regard to these molecular signatures.

Keywords: *Myelodysplastic syndrome, DNA micro-array, acute myeloid leukemia, stage progression, gene expression profile*

Introduction

Myelodysplastic syndrome (MDS) is an enigmatic disorder that is characterized by 2 clinical manifestations: ineffective hematopoiesis (cytopenia in peripheral blood despite hyper- or normal cellularity in bone marrow) and dysplastic morphology of blood cells [1]. MDS mostly affects the elderly, with an incidence of 15–50 cases per 100 000 people per year [2]. Clonality in multiple lineages of blood cells is found in individuals with MDS, suggesting that MDS is a clonal disorder of multi-potent stem cells in bone marrow [3].

An important aspect of MDS is that it comprises different clinical stages. According to the World Health Organization (WHO) classification of MDS [4], affected individuals whose bone marrow contains < 5% blasts are diagnosed with refractory anemia (RA), RA with ringed sideroblasts (RARS), refractory cytopenia with multi-lineage dysplasia (RCMD) or refractory cytopenia with multi-lineage dysplasia and ringed sideroblasts (RCMD-RS), whereas those whose bone marrow contains 5–9% or 10–19% blasts are diagnosed with RA with excess blasts (RAEB)-1 or RAEB-2, respectively. About 10–30% of MDS patients at the early stages (RA, RARS, RCMD or RCMD-RS) will eventually undergo stage progression to RAEB and, subsequently, to acute myeloid leukemia (AML).

Despite the relatively high incidence of MDS, its molecular pathogenesis is poorly understood (Figure 1). Gene mutations or other genomic alterations that might give rise to RA or RCMD remain to be identified and the ineffective hematopoiesis apparent in MDS patients remains to be characterized at the molecular biological level. It is also not known what triggers progression of early stages of MDS to advanced ones in some individuals but not others.

Cytopenia in peripheral blood is also found in patients with aplastic anemia (AA). Although the bone marrow of most individuals with AA is characterized by hypocellularity, the difference in marrow cellularity between patients with AA and those with RA or RCMD is not always clear. Anti-thymocyte globulin, a standard treatment for AA, is also effective in a sub-set of patients at the early

Correspondence: Hiroyuki Mano, MD, PhD, Division of Functional Genomics, Jichi Medical School, 3311-1 Yakushiji, Kawachigun, Tochigi 329-0498, Japan. Tel: +81-285-58-7449. Fax: +81-285-44-7322. E-mail: hmano@jichi.ac.jp

Received for publication 15 July 2005.

ISSN 1042-8194 print/ISSN 1029-2403 online © 2006 Taylor & Francis
DOI: 10.1080/10428190500264231

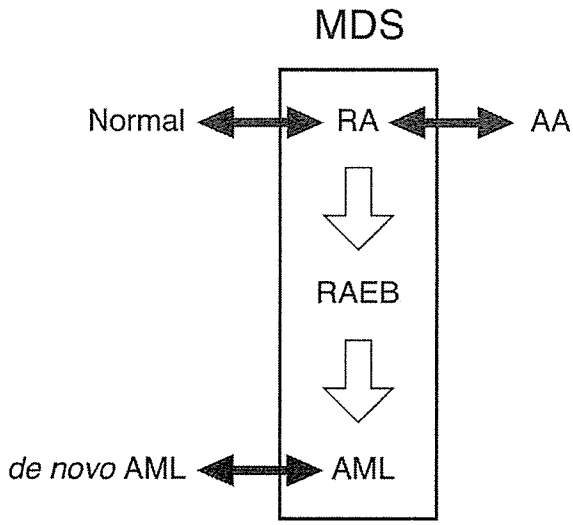

Figure 1. Stage progression of MDS and its relation to other conditions. Little is currently known of the molecular pathogenesis of MDS or of the similarities or differences between the early stages of MDS and aplastic anemia (AA) and between MDS-derived AML and *de novo* AML.

stages of MDS [5], further complicating the distinction between the two disorders.

It is widely believed that AML evolved from MDS has a poorer prognosis than does *de novo* AML, suggesting that the 2 clinical entities might be distinct. However, patients with MDS-associated AML are often older than are those with *de novo* AML and tend to possess karyotypes associated with high risk. It, thus, remains unclear whether the prognosis of *de novo* AML does indeed differ from that of MDS-associated AML if patient age and karyotype are matched.

The Human Genome Project is now close to completion, with 99% of DNA in euchromatin having been sequenced at 99.999% accuracy (http://www.ncbi.nlm.nih.gov/genome/guide/human/). Annotation of the human genome has revealed an unexpectedly small number (20 000–25 000) of protein-coding genes [6] compared with the numbers identified in the nematode (~19 000 genes) [7] and fruit fly (~14 000 genes) genomes [8]. The development of DNA micro-array analysis now allows simultaneous measurement of the level of expression of tens of thousands of genes in a given sample [9,10]. With this approach, it is thus possible to obtain a total gene expression profile or 'transcriptome' for each of the various stages of MDS and then to compare these profiles either among MDS stages or with those of the healthy condition, AA or *de novo* AML. Such analysis has the potential both to identify novel molecular markers for the differential diagnosis of MDS vs AA or *de novo* AML as well as to reveal genes that contribute to the pathogenesis of MDS. It might also be possible to determine whether MDS should be treated as a clinical entity distinct from AA or *de novo* AML.

It is important to bear in mind, however, that blood cells of different lineages and differentiation levels possess markedly different transcriptomes, even within the same individual. Any shift in cell composition in the specimens analysed will, thus, greatly influence the gene expression profile determined by micro-array studies [11].

Normal vs MDS

Several studies have attempted to compare transcriptomes between healthy individuals and patients with MDS in order to identify differences in gene expression. Pellagatti et al. [12] isolated RNA from differentiated granulocytes of 7 healthy individuals and 21 patients with MDS (17 with RA, 2 with RARS and 2 with RAEB according to the French-American-British (FAB) classification). The RNA was subjected to hybridization with a cDNA micro-array harboring probes corresponding to ~6000 human genes and the researchers identified 12 genes whose expression was frequently up-regulated (ratio of >2.0 in $\geqslant 9$ patients) or down-regulated (ratio of <0.5 in $\geqslant 10$ patients) in MDS. The relevance of these genes to the molecular diagnosis of MDS is unknown.

In contrast, Hofmann et al. [13] isolated $CD34^+$ progenitor fractions from the bone marrow of 4 healthy subjects and 11 MDS patients (7 low-risk and 4 high-risk according to the International Prognostic Scoring System [14]) for analysis with micro-arrays containing >12 000 human probe sets. They identified 161 genes whose expression was down-regulated (ratio of <0.2) in low-risk MDS patients compared with healthy subjects. They also detected 117 genes whose expression was up-regulated (ratio of >5) in MDS patients and 27 of these genes encoded regulators of hematopoiesis, including acute myeloid leukemia 1 (AML1), activating transcriptional factor 3 (ATF3), homeobox 7 (HOX7) and Delta-like homolog 1 (DLK1).

Chen et al. [15] also chose $CD34^+$ cells for comparison of transcriptomes between healthy controls and MDS patients, specifically those with monosomy 7 or trisomy 8. $CD34^+$ progenitor cells were purified from the bone marrow of 4 control subjects, 4 MDS patients with trisomy 8 and 2 MDS patients with monosomy 7 and RNA isolated from these cells was subjected to hybridization with the same type of arrays (Affymetrix GeneChip HGU95Av2) as those used by Hofmann et al. [13]. Comparison of gene expression profiles among the subjects revealed that genes important in immune function and inflammation were frequently over-expressed in the MDS patients

with trisomy 8, consistent with previous findings implicating autoimmune activity in such patients [16]. In contrast, genes important in cell growth were often down-regulated in the patients with monosomy 7. These findings, thus, suggested that gene expression profiles differ between MDS blasts with trisomy 8 and those with monosomy 7.

Stage progression in MDS

A substantial proportion of MDS patients in early stages of the disease, especially those with an unfavorable karyotype [17], undergoes progression to advanced stages or to AML. Given that currently available chemotherapeutic regimens for advanced MDS are of limited efficacy, it would be clinically advantageous to block stage progression in MDS. Alterations of several oncogenes and tumor suppressor genes have been implicated in the progression of MDS [18]. Activating mutations of *RAS* genes are thought to be the most prevalent (affecting 10 – 30% of cases) of such changes in MDS [19,20]. It remains unclear, however, whether *RAS* mutation occurs at the early or late stages of MDS. Inactivation of the p53 gene is also apparent in 5 – 10% of MDS patients [21]. Again, however, it is not known whether loss of p53 function is an early or late event during MDS progression. In addition, epigenetic silencing of the p15 gene and shortening of telomeres have been detected in bone marrow cells of MDS patients [22]. None of these gene alterations is specific to MDS and it is unclear which changes are the cause of MDS itself and which are associated with stage progression.

To obtain insight into the mechanism of stage progression of MDS, in their comparison of gene expression profiles among healthy controls, low-risk MDS patients and high-risk MDS patients, Hofmann et al. [13] applied the class membership prediction method to identify genes whose expression was linked to separation of the 3 classes. They identified 11 such genes (Table I) and a simple 2-way clustering analysis of the study subjects based on the expression patterns of these 11 genes clearly separated the 3 major classes. Furthermore, a similar clustering analysis of a second set of subjects ($n = 8$) also separated individuals with high-risk MDS from those with low-risk MDS. Although the number of study subjects was small, these data support the notion that each stage of MDS has a characteristic gene expression profile or 'molecular signature'.

Table I. Genes used for class separation of healthy individuals and low- or high-risk patients with MDS [13].

Gene symbol	Accession number	Chromosomal position
TACSTD2	J04152	1p32
UQCRC1	L16842	3p21.3
TNNC1	M37984	3p21.3
KDELR	M88458	7p
CLC	L01664	19q13.1
H-PKL	M5422	7
RGS19	Z91809	
ATF3	L19871	1
FARP1	AI701049	
GNG7	AW051450	
TPD52L2	U44429	6q22-q23

CD133 (also known as AC133) is a cell surface protein that is expressed exclusively on $CD34^+$ $CD38^-$ hematopoietic stem cells (HSCs) [23,24]. Many AML blasts also express CD133 [25], indicating that the differentiation of these cells is blocked at a highly immature stage. The existence of 'cancer stem cells' for AML and solid tumors has been recently demonstrated and such cells express CD133 in brain tumors [26], as do $CD34^+CD38^-$ cells in AML [27]. Analysis of $CD133^+$ HSC-like fractions among MDS patients may, thus, reveal the character of 'MDS stem cells'. Analysis of such fractions also has the advantage of eliminating from micro-array data the influence of variation in cell composition of bone marrow, which is especially important given that different stages of MDS are characterized by different numbers of immature blasts within marrow.

Ueda et al. [28] performed micro-array analysis with $CD133^+$ cells isolated from the bone marrow of 2 healthy individuals, 11 patients with RA, 5 patients with RAEB and 14 patients with MDS-associated AML. Comparison of the gene expression profiles among the different stages of MDS led to the identification of 11 late stage (RAEB, MDS-associated AML)-specific genes and 6 early stage (healthy controls, RA)-specific genes. The latter set of genes included that for PIASy, which catalyses sumoylation of substrate proteins [29]. Loss of expression of the PIASy gene in advanced MDS suggested that the encoded protein might possess anti-tumor activity. Consistent with this notion, forced expression of PIASy in a mouse myeloid cell line resulted in rapid induction of apoptosis when the cells were cultured in the presence of granulocyte colony-stimulating factor (G-CSF) (Figure 2) [28]. These results suggest that PIASy functions to restrain cell growth and that loss of its expression may facilitate stage progression in MDS. Loss of PIASy expression has also been implicated in stage progression of chronic myeloid leukemia [30].

MDS-derived AML vs *de novo* AML

Although dysplastic morphology of blood cells is a hallmark of MDS and MDS-derived AML, such

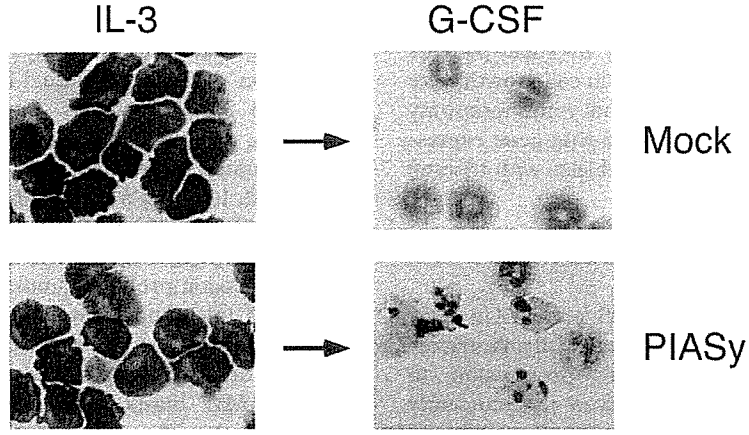

Figure 2. Induction of apoptosis by PIASy. 32Dcl3 cells were infected with a control retrovirus (Mock) or with a virus that encodes human PIASy and were then either maintained in the presence of interleukin-3 (IL-3) or exposed to G-CSF. Whereas control cells incubated with G-CSF underwent gradual differentiation into granulocyte-like cells, those expressing PIASy died rapidly by apoptosis on exposure to G-CSF. Reproduced with permission from Ueda et al. [28].

dysplasia is also apparent in the blood cells of some healthy older individuals. It is, thus, sometimes difficult to differentiate *de novo* AML from MDS-derived AML in the elderly, especially in the absence of a clinical history of the patient. Further complicating the issue, some younger patients with *de novo* AML also manifest blood cell dysplasia [31,32].

To provide insight into the differences or similarities between these clinical classes, Oshima et al. purified $CD133^+$ cells from the bone marrow both of 10 patients with *de novo* AML of the M2 sub-type, according to the FAB classification [33] and of 10 patients with MDS-derived AML corresponding to the M2 sub-type [34]. They then subjected RNA from these cells to micro-array analysis with HGU95Av2 micro-arrays. Comparison of samples matched for FAB sub-type was performed to minimize the influence of the differentiation ability of the blasts on gene expression profile; any differences in gene expression identified with this approach would, thus, be expected to be related with a high probability to the difference in the nature of MDS-derived AML from that of *de novo* AML.

Statistical analysis of the resulting expression data (Welch's ANOVA, $p < 0.01$; effect size of $\geqslant 5.0$ units) identified a total of 57 probe sets corresponding to genes whose expression was associated with diagnosis. However, a simple 2-way clustering analysis of the subjects based on the expression pattern for these probe sets failed to separate them into diagnosis-related sub-groups. To visualize the similarity or difference between the 2 classes, the researchers applied correspondence analysis, a method for the decomposition of multi-dimensional data [35]. This approach allows not only a low-dimensional projection of the expression profiles of numerous genes but also measurement both of the contribution of each gene to a given extracted dimension and of the contribution of each extracted dimension to the total complexity.

Correspondence analysis of the expression data for the 57 probe sets reduced the complexity from 57 to 3 dimensions. The specimens were then projected into a virtual space on the basis of their calculated 3-dimensional (3D) co-ordinates (Figure 3(*a*)). Most subjects with *de novo* AML were localized in a region of the space distinct from that occupied by those with MDS-derived AML. However, 2 individuals with *de novo* AML localized with those with MDS-derived AML. These observations indicated that the transcriptome of MDS-derived AML is distinct from that of *de novo* AML, but that current clinical diagnosis does not completely correlate with the difference in transcriptomes.

A similar analysis was performed on a larger scale by Tsutsumi et al. [36]. These researchers isolated $CD133^+$ cells from the bone marrow of patients with MDS-derived AML ($n = 11$), with *de novo* AML without dysplasia ($n = 15$), with *de novo* AML with multi-lineage dysplasia ($n = 11$) [32,37] or with therapy-related AML ($n = 2$). The study subjects were not limited to a specific FAB sub-type, however. The transcriptomes of these clinical classes were compared with the use of HGU95Av2 arrays. Comparison of *de novo* AML without dysplasia and MDS-derived AML led to the identification of 30 probe sets corresponding to genes whose expression was related to diagnosis. Correspondence analysis